Positive Definite Kernels for Partitions

JYRKO CORREA-MORRIS
Mathemtics and Natural Sciences Department
Miami Dade College, Padron Campus
627 SW 27 Ave, Miami, Florida
USA

Abstract: This paper presents a comprehensive exploration of various families of positive definite kernels for comparing partitions. It not only reviews existing examples from the literature but also introduces novel classes of positive definite kernels. These new classes include kernels based on agreement and ones designed using the concept of hidden variables. The study also focuses on assessing the compatibility of these kernels with structural properties that capture the intrinsic notion of proximity between partitions. Notably, agreement-based kernels are demonstrated to align well with this notion. Moreover, the paper provides two generic procedures for designing hidden-feature-based kernels that also adhere to the specified structural properties.

Key-Words: clustering, definite positive kernels, lattice of partitions

Received: May 11, 2023. Revised: August 12, 2023. Accepted: September 23, 2023. Published: October 9, 2023.

1 Introduction

1.1 Overview.

Since the turn of the century, there has been a significant increase in interest in exploring existing clustering methods, [1], [2], [3], within the field of pattern recognition. This surge in interest can be traced back to J. Kleinberg's seminal paper on an impossibility theorem for clustering, [4], which has served as a catalyst for other formal studies, [5], [6], [7].

The exploration of measures for comparing partitions has also been influenced by this growing interest. Evidently, notable contributions have been made, [8], [9], [10], [11]. There seems to be a consensus regarding the existence of fundamental principles that govern the natural proximity of partitions, particularly when dealing with structural data. For instance, when comparing graphs or trees, the editing distance, despite its computational complexity, is often considered the natural proximity measure, and proposed measures are evaluated based on their compatibility with this intrinsic paradigm, [12]. In this paper, we focus on studying an almost unexplored class of measures: that of positive definite kernels for comparing partitions. A similarity measure $k : X \times X \to \mathbb{R}$ is said to be a positive definite kernel when the dataset X can be embedded into a high-dimensional Euclidean (possibly an infinite-dimensional Hilbert) space $\mathcal{H}$ by means of a map φ such that the similarity $k(x, x')$ between the original data x and x' matches the dot product $\langle\varphi(x), \varphi(x')\rangle_{\mathcal{H}}$ of the vectors $\varphi(x)$ and $\varphi(x')$ in $\mathcal{H}$, [13]. The map $\varphi : X \to \mathcal{H}$, the space $\mathcal{H}$, and the vectors $\varphi(x)$, $x \in X$, are referred to as the *feature map*, *feature space*, and *feature vectors*, respectively.

Kernel-based methods have emerged as powerful tools for efficiently handling complex data such as strings, graphs, images, documents, and multimodal data, [14]. The key insight is that by computing similarities as dot products in a high-dimensional feature space $\mathcal{H}$ without explicitly visiting it, linear models can be effectively applied in non-linear settings, [13]. To achieve this, a nonlinear feature map φ is utilized to map the data into the high-dimensional space $\mathcal{H}$, where linear models can be employed. This approach allows for the utilization of additional structures of the feature space (e.g., vector space structure, topological structure) that may not be present in the original data space. As a result, operations like arithmetical operations on feature vectors can be extended to graphs, histograms, and partitions, providing greater flexibility in handling diverse data types.

However, the success of kernel-based methods largely depends on the choice of the kernel measure, which should align with the specific application and data characteristics.

In the context of clustering, the term "clustering kernel" is often associated with the kernel K-means method, [15]. This method involves mapping the original data X into the feature space $\mathcal{H}$, enabling K-means to be performed in the transformed space using an implicit nonlinear map φ. A challenge with this approach is that the explicit knowledge of the feature map is not available, making the updating of centroids cumbersome. Nonetheless, kernel K-means has demonstrated empirical success in identifying complex clusters that are non-linearly separable in the original data space. Furthermore, spectral clustering, which leverages eigenvectors of a certain matrix computed from the similarity matrix, has been shown to be a special case of kernel K-means, [16]. Although variants of this method exist (e.g.,

[17], [18], [19], [20]), their contributions to kernel construction and design are somewhat limited.

In contrast, a cluster ensemble algorithm (see, for instance, [21]) that utilized positive definite kernels to compare partitions and construct average-based consensus functions was reported, [22]. The authors introduced "Set significance-based kernels" for this purpose. Recognizing that maximizing consensus functions can be computationally expensive, even with simple measures like the Rand Index, [23], the authors transformed the consensus problem from the space of partitions to the feature space, which is a high-dimensional Euclidean space $\mathcal{H}$. This shift allowed for the use of well-known and easily computable solutions to the problem of maximizing average-based consensus functions in the feature space. The article presented a novel approach to tackle the cluster ensemble problem, reducing the task of maximizing consensus functions to a kernel preimage problem. In other words, the consensus solution, computed exactly in the feature space, needed to be brought back to the space of partitions. However, [22] made limited progress in solving the preimage problem, as it heavily depended on the specific kernel used. Recent contributions in the field have shown significant advancements towards achieving this goal for specific kernels. For instance, [24] focused on consensus functions defined from the Rand index, while [25] concentrated on prototype-based kernels. Nevertheless, this area of research remains rich with several unanswered questions.

Driven by these inquiries, this paper presents several classes of positive kernels for partitions and investigates their behavior by analyzing structural properties that emulate the inherent proximity between partitions.

1.2 The natural notion of proximity between partitions

Let S be a finite dataset. A partition p of S is a collection of nonempty subsets $\{C_1, C_2, \dots, C_s\}$ of S, called the *clusters* of p, such that $S = \cup_{i=1}^{s} C_i$ and $C_i \cap C_j = \emptyset$ if $i \neq j$. The set of all partitions of S (henceforth denoted by $\mathbb{P}_S$) is endowed with the refinement of partitions. The partition p is said to *refine* the partition p' if, and only if, every cluster of p' is the union of some clusters of p. In particular, every cluster of p is contained in some cluster of p'. The notation $\mathrm{p} \preceq \mathrm{p}'$ means p refines p'.

Also, $\mathbb{P}_S$ is endowed with two operations: the *meet* $\wedge$ and the *join* $\vee$, which assign to any two given partitions p and p' the coarsest partition $\mathrm{p} \wedge \mathrm{p}'$ that refines p and p' and the finer partition simultaneously refined by p and p', respectively. Two elements $x, x' \in S$ are in the same cluster of $\mathrm{p} \wedge \mathrm{p}'$ when they are in the same cluster of p and in the same cluster of p'. Two elements $x, x' \in S$ are placed in the same cluster of $\mathrm{p} \vee \mathrm{p}'$ when there is a sequence $x = x_{i_1}, x_{i_2}, \dots, x_{i_k} = x'$ such that, for all $j \in \{1, 2, \dots, k-1\}$, x_{i_j} and $x_{i_{j+1}}$ are either placed in the same cluster of p or in the same cluster of p'.

$\mathbb{P}_S$ can be spatially organized into layers of partitions such that only partitions with the same number of clusters are placed in the same layer: (1) Place the all-singleton partition in the bottom layer. (2) A partition p' *covers* another partition p when p refines p' and there is no partition p'' refined by p that also refine p'. In the kth layer, place all partitions that *cover* some partition placed in the $(k-1)$th layer. (3) A line segment is drawn between two partitions p and p' if one of them covers the other. (The resulting diagram can be seen here: https://blogs.ams.org/visualinsight/2015/06/15/lattice-of-partitions/)

The underlying notion of proximity among partitions is associated to this spatial organization. Indeed, since the Hasse diagram is a connected graph, we can reach one partition from another by exclusively traveling along the edges of this diagram, the shorter the path connecting two partitions, the greater the similarity between these partitions. To see some of the characteristic features of such a proximity, let us start by considering an ascending path starting at a partition p. Note that every partition in this path is not only refined by p, but by all the partitions that precede it in the path. Thus, if p' and p'' are two partitions in this path such that p' is reached before p'' when traveling this path from p, we have that p refines both p' and p'', p' refines p'', and obviously, p is closer to p' than to p''. Since three partitions $\mathrm{p} \preceq \mathrm{p}' \preceq \mathrm{p}''$ are always part of an ascending path starting at p, such distinguishing quality of the natural proximity can be formally stated as follows:

Bottom-Up Collinearity: If $\mathrm{p} \preceq \mathrm{p}' \preceq \mathrm{p}''$, then p is more similar to p' than to p''. Thus, it is expected that any similarity measure k for comparing partitions (in particular, positive kernels) satisfies: $k(\mathrm{p}, \mathrm{p}') \geq k(\mathrm{p}, \mathrm{p}'')$, provided that $\mathrm{p} \preceq \mathrm{p}' \preceq \mathrm{p}''$.

Analogously, we can take instead a descending path starting at a partition p''. All partitions that we find along the way not only refine p'' but all the partitions that appear before it in this path. Consequently, if the partition p is farther from p'' in this path than another partition p', then p'' is refined by both p' and p, and p' is refined by p. This distinguishing property can be formally formulated as follows:

Top-Down Collinearity: If $\mathrm{p} \preceq \mathrm{p}' \preceq \mathrm{p}''$, then p'' is more similar to p' than to p. Accordingly, it is expected that any similarity measure k for comparing partitions (in particular, positive kernels) satisfies: $k(\mathrm{p}'', \mathrm{p}') \geq k(\mathrm{p}'', \mathrm{p})$, as long as $\mathrm{p} \preceq \mathrm{p}' \preceq \mathrm{p}''$.

The two properties above only involve partitions either in an ascending path or a descending path of the diagram that describes the natural spatial organization of partitions. As we saw above, these conditions require for any two considered partitions that one refines the other. How about partitions p and p' such that neither of them refines the other? Note that, in such a case, any path from one of these partitions to the other either ascends first and then it descends, or it descends first and then it ascends. The partition at which this path turns back happens to be the join $p \vee p'$ of p and p' and the meet $p \wedge p'$ of the partitions p and p', respectively. Thus, the similarity between two partitions p and p' is at most equal to the highest similarity value from among the similarity values between the partitions p and $p \vee p'$ and between the partitions p and $p \wedge p'$. This leads us to the following property.

Meet-Join Predominance: The similarity between two partitions p and p' never exceeds the maximum value from among the similarity between p and $p \vee p'$ and the similarity between p and $p \wedge p'$. Thus, it is expected that, for any partitions p and p', $k(p, p') \leq \max\{k(p, p \vee p'), k(p, p \wedge p')\}$. When for any two partitions, the previous inequality always holds with the meet (resp. with the join), we name this property *Meet predominance* (resp. *Join predominance*).

1.3 Contributions and organization

The main contributions of this paper are: (1) The class of agreement-based kernels is introduced and formally studied. In particular, the Rank kernel and consensus-based kernels are presented as paradigms of this family. (2) Kernels designed from a set of hidden features are investigated. Special attention is dedicated to significance-set-based kernels and diameter-based kernels. (3) The compliance of positive definite kernels with the intrinsic notion of proximity between partition is examined in detail. Analogous properties have been previously employed to characterize metrics used for comparing partitions, [11]. In that work, these properties were applied to study the behavior of the metrics. Additionally, the compatibility between these properties and submodularity has been investigated, [26].

Aside from Introduction, the paper is organized in five additional sections. In Section 2, agreement-based kernels are introduced and their theoretical foundation is established. Section 3 introduces and studies the class of kernels designed from hidden features. Particularly, two main procedures for the purpose of designing hidden features whose associated kernel is in compliance with intrinsic notion of proximity between partitions are presented. Section 4 shows that prototype-based kernels do not respect the natural notion of proximity between partitions. Section 5 is devoted to the conclusions and future work, while the proofs of the main results are included in the appendix.

2 Agreement-based Kernels

In this section, J denotes the set of all possible unordered pairs of elements in S while, for a given partition p of S, $J(\mathrm{p})$ denotes the set of those pairs in J that lie in the same cluster of p. For two partitions p and p' of S, there are the sets: $N_{00} = J(\mathrm{p}) \cap J(\mathrm{p}')$, $N_{01} = J(\mathrm{p}) \cap J(\mathrm{p}')^C$, $N_{10} = J(\mathrm{p})^C \cap J(\mathrm{p}')$, and $N_{11} = J(\mathrm{p})^C \cap J(\mathrm{p}')^C$, where the superscript C indicates the complement of the set. The sizes of these sets are denoted by $\mathfrak{n}_{00}$, $\mathfrak{n}_{01}$, $\mathfrak{n}_{10}$, and $\mathfrak{n}_{11}$, respectively.

If a pair of objects is either in N_{00} or N_{11} (resp. N_{10} or N_{10}), then both partitions p and p' agree with (resp. differ in) regard to whether these objects should be placed in the same cluster or not. Thus, $\mathfrak{n}_{00} + \mathfrak{n}_{11}$ (resp. $\mathfrak{n}_{01} + \mathfrak{n}_{10}$) represents the total number of agreements (resp. disagreements) between p and p'. Note also that every pair belong to exactly one of these sets and therefore $\binom{n}{2} = \mathfrak{n}_{00} + \mathfrak{n}_{01} + \mathfrak{n}_{11} + \mathfrak{n}_{11}$. The following identities hold:

$$\mathfrak{n}_{00} = \sum_{i=1}^{s}\sum_{j=1}^{r}\binom{n_{ij}}{2};$$
$$\mathfrak{n}_{01} = \sum_{i=1}^{s}\binom{n_i}{2} - \sum_{i=1}^{s}\sum_{j=1}^{r}\binom{n_{ij}}{2};$$
$$\mathfrak{n}_{10} = \sum_{j=1}^{r}\binom{n'_j}{2} - \sum_{i=1}^{p}\sum_{j=1}^{q}\binom{n_{ij}}{2};$$
$$\mathfrak{n}_{11} = \binom{n}{2} - \mathfrak{n}_{00} - \mathfrak{n}_{01} - \mathfrak{n}_{10};$$

where n_i, n'_j, and n_{ij} denote the sizes of the ith cluster of p ($1 \leq i \leq s$), the jth cluster of p' ($1 \leq j \leq r$), and the overlapping between the ith cluster of p and the jth cluster of p', respectively.

To define an agreement-based kernel, we consider non-negative functions $\omega, \nu : J \to \mathbb{R}^+$ (not necessarily different), which associate a non-negative weight to each pair of objects $\{x, x'\}$. The agreement-based kernel corresponding to the weight functions ω and ν is defined by:

$$k_{\omega\nu}(\mathrm{p}, \mathrm{p}') := \sum_{\{x,x'\} \in N_{00}} \omega(x, x') + \sum_{\{x,x'\} \in N_{11}} \nu(x, x'). \tag{1}$$

Proposition 1. *Any agreement kernel $k_{\omega\nu}$ is a positive definite kernel.*

Agreement-based kernels are compatible with the intrinsic notion of proximity on the lattice of partitions of S.

Proposition 2. *Any agreement kernel $k_{\omega\nu}$ satisfies Collinearity, Dual Collinearity, and Meet Predominance. However, these kernels fail to satisfy Join Predominance.*

2.1 Rand Index

The Rand Kernel is perhaps the most known agreement-based kernel. It quantifies the similarity between two partitions p and p′ by counting the total number of agreements, all pairs being considered to have the same weight (i.e., for every pair $\{x, x'\} \in J$, $\omega(\{x, x'\}) = \nu(\{x, x'\}) = 1$). Thus,

$$RK(\mathrm{p}, \mathrm{p}') = \mathfrak{n}_{11} + \mathfrak{n}_{00}. \tag{2}$$

The original Rand index, published in 1971 by W. M. Rand, is given by $RI(\mathrm{p}, \mathrm{p}') = \frac{\mathfrak{n}_{11}+\mathfrak{n}_{00}}{\mathfrak{n}_{11}+\mathfrak{n}_{10}+\mathfrak{n}_{01}+\mathfrak{n}_{00}}$, and hence $RI(\mathrm{p}, \mathrm{p}') = \frac{RK(\mathrm{p},\mathrm{p}')}{\binom{n}{2}}$. On the other hand, we can normalize any kernel K by setting $NK(\mathrm{p}, \mathrm{p}') = \frac{K(\mathrm{p},\mathrm{p}')}{\sqrt{K(\mathrm{p},\mathrm{p})K(\mathrm{p}',\mathrm{p}')}}$. Since $RK(\mathrm{p}, \mathrm{p}) = RK(\mathrm{p}', \mathrm{p}') = \binom{n}{2}$, we get $NRI(\mathrm{p}, \mathrm{p}') = \frac{RK(\mathrm{p},\mathrm{p}')}{\binom{n}{2}} = RI(\mathrm{p}, \mathrm{p}')$, which proves that the original Rand index concurs with the normalized Rand kernel.

To show that agreement-based kernels fail in general to satisfy Join Predominance, let us consider the partitions $\mathrm{p} = \{\{x_1, x_2\}, \{x_3, x_4\}\}$ and $\mathrm{p}' = \{\{x_1, x_2, x_3\}, \{x_4\}\}$ of $S = \{x_1, x_2, x_3, x_4\}$. Note that $\mathrm{p} \vee \mathrm{p}' = \{\{x_1, x_2, x_3, x_4\}\}$ and that, for p and p′, $N_{00} = \{\{x_1, x_2\}\}$, $N_{01} = \{\{x_1, x_3\}, \{x_2, x_3\}\}$, $N_{10} = \{\{x_3, x_4\}\}$, and $N_{11} = \{\{x_1, x_4\}, \{x_2, x_4\}\}$; while for p and $\mathrm{p} \vee \mathrm{p}'$, $N_{00} = \{\{x_1, x_2\}, \{x_3, x_4\}\}$, $N_{01} = \{\{x_1, x_3\}, \{x_1, x_4\}, \{x_1, x_3\}, \{x_2, x_3\}, \{x_2, x_3\}\}$ (N_{10} and N_{11} are empty). This gives $RK(\mathrm{p}, \mathrm{p}') = 3$ and $RK(\mathrm{p}, \mathrm{p} \vee \mathrm{p}') = 2$, which means that p is more similar to p′ than to $\mathrm{p} \vee \mathrm{p}'$.

To end, note that, closely related to the Rand kernel is its counterpart: the well-known Symmetric Difference metric:

$$SDM(\mathrm{p}, \mathrm{p}') = \mathfrak{n}_{10} + \mathfrak{n}_{01}; \tag{3}$$

defined as the total number of disagreements. Thus, given partitions p and p′, the equality $RK(\mathrm{p}, \mathrm{p}') + SDM(\mathrm{p}, \mathrm{p}') = \binom{n}{2}$ holds. Any positive definite kernel k has an associated metric M_k, which is defined by $M_k(\mathrm{p}, \mathrm{p}') = k(\mathrm{p}, \mathrm{p}) + k(\mathrm{p}', \mathrm{p}') - 2\,k(\mathrm{p}, \mathrm{p}')$. SDM is (up to a constant factor) the metric associated to RK:

$$\begin{aligned} M_{RK}(\mathrm{p}, \mathrm{p}') &= RK(\mathrm{p}, \mathrm{p}) + RK(\mathrm{p}', \mathrm{p}') - 2\,RK(\mathrm{p}, \mathrm{p}') \\ &= 2\binom{n}{2} - 2\,[\mathfrak{n}_{11} + \mathfrak{n}_{00}] \\ &= 2\,[\mathfrak{n}_{01} + \mathfrak{n}_{10}] \\ &= 2\,SDM(\mathrm{p}, \mathrm{p}'). \end{aligned}$$

2.2 Consensus Kernels

The motivation for consensus kernels lies in the ostensible benefits of incorporating some prior knowledge. The intuition here is to use the co-association matrix associated to a given ensemble $\mathcal{E}$ to determine the weights of the pairs. Formally, given an ensemble $\mathcal{E} = \{\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_m\}$ of partitions of S, the associated *co-association matrix* is the $n \times n$ matrix $A_{\mathcal{E}} = (a_{ij})_{i,j=1}^{m}$ whose (i, j)-entry carries the ratio between the number of partitions in $\mathcal{E}$ for which the pair $\{x_i, x_j\}$ is placed in the same cluster and the total number the partitions in the ensemble.

Defining the weight functions ω and ν from the matrix $A_{\mathcal{E}}$ by $\omega(\{x_i, x_j\}) = a_{ij}$ and $\nu(\{x_i, x_j\}) = 1 - a_{ij}$, we define the consensus kernels associated to the ensemble $\mathcal{E}$ by:

$$k_{\mathcal{E}}(\mathrm{p}, \mathrm{p}') = \sum_{\{x_i,x_j\}\in N_{00}} a_{ij} + \sum_{\{x_i,x_j\}\in N_{11}} (1-a_{ij}). \tag{4}$$

3 Another family of positive definite kernel: Hidden features

Hidden features is a term used to designate data features that are not actually observed but rather determined by or inferred from the observed features. Thus, hidden features are associated to functions $f_i : X \to \mathbb{R}$ that assign every datum, in our case a partition p, to the measurement $f_i(\mathrm{p})$ of the ith hidden feature. Every partition is then represented by the vector $\varphi(\mathrm{p}) = (f_1(\mathrm{p}), f_2(\mathrm{p}), \ldots, f_i(\mathrm{p}), \ldots)$. Given a finite or enumerable set of hidden features, a positive defined kernel for comparing partitions can be defined by setting $k(\mathrm{p}, \mathrm{p}') = \sum_i f_i(\mathrm{p}) \cdot f_i(\mathrm{p}')$.

In this section, we seek sufficient conditions for these kernels to be in compliance with the natural proximity between partition. In this regard, Theorem 1 provides a procedure that enables us to construct many examples of such kernels.

Theorem 1. *Let f_i, $i \in I$ (finite set), be hidden features for the partitions of the dataset S. Suppose that the following conditions hold:*

(i) *For all* $\mathrm{p} \in \mathbb{P}_S$, $f_i(\mathrm{p}) \geq 0$.

(ii) $\mathrm{p} \prec \mathrm{p}'$ *and* $f_i(\mathrm{p}) > 0$ *imply* $f_i(\mathrm{p}) \geq f_i(\mathrm{p}') > 0$. *Moreover, the equality holds for all hidden features f_i if, and only if,* p *and* p′ *are the same partition.*

(iii) $\mathrm{p} \prec \mathrm{p}'$ *implies*

$$\sum_{i, f_i(\mathrm{p})>0} f_i(\mathrm{p}) - \sum_{i, f_i(\mathrm{p})>0} f_i(\mathrm{p}') < \sum_{i, f_i(\mathrm{p})=0} f_i(\mathrm{p}'). \tag{5}$$

(iv) *For all* $\mathrm{p}, \mathrm{p}' \in \mathbb{P}_S$, *if both* $f_i(\mathrm{p})$ *and* $f_i(\mathrm{p}')$ *are positive, then* $f_i(\mathrm{p} \wedge \mathrm{p}')$ *is positive as well.*

Then the kernel $k(\mathrm{p}, \mathrm{p}') = \sum_{i\in I} f_i(\mathrm{p}) \cdot f_i(\mathrm{p}')$ *satisfies Bottom-Up Collinearity, Top-Down Collinearity, and Meet Predominance.*

Condition (i) establishes that all feature vectors $\varphi(\mathrm{p})$ have all their components nonnegative, which amounts to saying that the feature map φ embeds $\mathbb{P}_S$ into the first closed orthant of the p-dimensional Euclidean space, where p is the number of hidden features. Condition (ii) asserts that, if the measurement of a hidden feature f_i at the partition p is positive, then it decreases along any ascending path starting at p without vanishing. A word of caution is necessary here. Condition (ii) must not be mistaken with that which asserts that f_i decreases along any ascending path without vanishing. Condition (ii) allows that $f_i(\mathrm{p}) = 0$ and $f_i(\mathrm{p}') > 0$, for some p' refined by p. Additionally, Condition (iii) establishes that the number of hidden features that vanish at p sufficiently exceeds the number of such features that vanishes at p'. To get a better insight into (iii), let $\varphi(\mathrm{p})_+$ and $\varphi(\mathrm{p}')_+$ denote the vectors obtained from $\varphi(\mathrm{p})$ and $\varphi(\mathrm{p}')$ by considering only those hidden variables that are positive when measured on p, and let $\varphi(\mathrm{p}')_0$ be the vector obtained from $\varphi(\mathrm{p}')$ by considering only those hidden variables that vanish when measured on p. In terms of the 1-norm of Euclidean spaces, (ii) implies that $||\varphi(\mathrm{p})||_1 = ||\varphi(\mathrm{p})_+||_1 > ||\varphi(\mathrm{p}')_+||_1$ while (iii) can be restated as $||\varphi(\mathrm{p})_+||_1 - ||\varphi(\mathrm{p}')_+||_1 < ||\varphi(\mathrm{p}')_0||_1$, which forces $||\varphi(\mathrm{p})||_1 < ||\varphi(\mathrm{p}')||_1$. Finally, Condition (iv) prevents f_i from vanishing at the meet $\mathrm{p} \wedge \mathrm{p}'$ of two partitions p and p' if f_i returns a positive measurement for both partitions p and p'.

3.1 Set significance based kernels

The set-significance-based kernels were introduce by Vega-Pons et. al. [22]. In this case, the hidden variables are the subsets of S. Given a partition p, how significant with respect to p every subset $\mathfrak{s} \subseteq S$ it is determined. Assuming that for a partition p the most significant subsets of S are its clusters, the closer to be a cluster of p the subset $\mathfrak{s}$ is, the more significant for p. As an attempt to capture this essence, the following definition is given. A function $\mu : 2^S \times \mathbb{P}_S \to \mathbb{R}_{\geq 0}$ is said to be a measure of set significance if *for all clusters* c *and* c' *of* p *and all same-size subsets* $\mathfrak{s}_1$ *and* $\mathfrak{s}_2$ *of* S *with* $\mathfrak{s}_1 \subseteq \mathrm{c}$ *and* $\mathfrak{s}_2 \subseteq \mathrm{c}'$, $\mu(\mathfrak{s}_1, \mathrm{p}) < \mu(\mathfrak{s}_2, \mathrm{p})$ *if, and only if,* c *is bigger than* c'; *and* $\mu(\mathfrak{s}_1, \mathrm{p}) = \mu(\mathfrak{s}_2, \mathrm{p})$ *if, and only if,* c *and* c' *have the same size.*

Listing the subsets $\mathfrak{s}_1, \mathfrak{s}_2, \ldots, \mathfrak{s}_{2^n}$ of S in a predetermined order, the ith hidden feature f_i measures the significance of ith subset $\mathfrak{s}_i$ of S through the formula $f_i(\mathrm{p}) = \mu(\mathfrak{s}_i, \mathrm{p})$.

In [22], the authors used the measure of significance:

$$\mu(\mathfrak{s}_i, \mathrm{p}) = \begin{cases} \frac{|\mathfrak{s}_i|}{|\mathrm{c}|}, & \text{if } \mathfrak{s}_i \subseteq \mathrm{c} \text{ for some } \mathrm{c} \in \mathrm{p}; \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where the bars | | indicate the size of the set.

The kernel k_μ is a typical example of a kernel constructed from the procedure described in Theorem 1.

Proposition 3. $k_\mu(\mathrm{p}, \mathrm{p}') = \frac{1}{4} \sum_{i=1}^{s} \sum_{j=1}^{r} 2^{n_{ij}} \cdot \left(\frac{n_{ij}^2 + n_{ij}}{n_i n_j'}\right)$. *Furthermore,* k_μ *satisfies Bottom-Up Collinearity, Top-Down Collinearity, and Meet Predominance.*

3.2 Expanding on Theorem 1

Theorem 1 provides a suitable procedure for the construction of clustering kernels that are consistent with the intrinsic notion of proximity between partition. However, this procedure hinges on sufficient conditions that not all such kernels are supposed to meet. This section gives an alternative procedure that also ensures the compliance of hidden-feature-based kernels with the aforementioned structural properties.

The idea consists of superseding conditions $(ii) - (iv)$ with some analogous ones that are applicable to families of hidden features that fail to satisfy those required by Theorem 1.

Theorem 2. *Let* f_i, $i \in I$ *(finite set), be hidden features for the partitions of the dataset* S. *Suppose that the following conditions hold:*

(i) *For all* $\mathrm{p} \in \mathbb{P}_S$, $f_i(\mathrm{p}) \geq 0$.

(v) $\mathrm{p} \prec \mathrm{p}'$ *and* $f_i(\mathrm{p}') > 0$ *imply* $f_i(\mathrm{p}') \geq f_i(\mathrm{p}) > 0$.

(vi) $\mathrm{p} \prec \mathrm{p}'$ *implies*

$$\sum_{i, f_i(\mathrm{p}')>0} f_i(\mathrm{p}') - \sum_{i, f_i(\mathrm{p}')>0} f_i(\mathrm{p}) < \sum_{i, f_i(\mathrm{p}')=0} f_i(\mathrm{p}). \tag{7}$$

Then the kernel $k(\mathrm{p}, \mathrm{p}') = \sum_{i\in I} f_i(\mathrm{p}) \cdot f_i(\mathrm{p}')$ *satisfies Bottom-Up Collinearity, Top-Down Collinearity, and Meet Predominance.*

Now we introduce a novel class of kernels for partitions whose conformity with the intrinsic proximity between partition can be justified by Theorem 2: diameter-based kernels.

Let us consider a similarity measure $\sigma : S \times S \to [0, 1]$ (note that σ does not quantify the similarity between partitions, but between the underlying data). By a similarity measure here we mean a symmetric function (i.e., $\sigma(x,x') = \sigma(x',x)$) such that every datum maximizes the similarity to itself (i.e., $\sigma(x,x') \leq \sigma(x,x)$ for all $x,x' \in S$). Given a subset $\mathfrak{s}$ of S, the diameter of $\mathfrak{s}$ with respect to σ is $\text{diam}(\mathfrak{s}) = \max\{\sigma(x,x') : x,x' \in \mathfrak{s}, x \neq x'\}$.

Like set-significance-based kernels, diameter-based kernels k_σ are designed using the subsets $\mathfrak{s}$ of S (sorted in a list) as the hidden variables. Denoting by $c_i(\mathrm{p})$ the total number of clusters of p contained in the subset $\mathfrak{s}_i$ and by and $\nu_i(\mathrm{p})$ the function indicator of non-singleton clusters of p within $\mathfrak{s}_i$, having the value 1 if there are singleton clusters of p within $\mathfrak{s}_i$ and the value 0 otherwise, the ith hidden feature is defined by:

$$f_i(\mathrm{p}) = \begin{cases} 1 + \frac{\nu_i(\mathrm{p})\cdot\text{diam}(\mathfrak{s}_i)}{c_i(\mathrm{p})}, & \text{if } c_i(\mathrm{p}) > 0; \\ 0, & \text{if } c_i(\mathrm{p}) = 0. \end{cases}$$

Proposition 4. *k_σ satisfies Bottom-Up Collinearity, Top-Down Collinearity, and Meet Predominance.*

4 A well-known family of not well-behaved kernels

We conclude with a brief note regarding the prototype-based kernels. These kernels are designed from a set of prototypes $\mathcal{P} := \{\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_m\} \subseteq \mathbb{P}_S$ and a dissimilarity measure $d : \mathbb{P}_S \times \mathbb{P}_S \to \mathbb{R}$ (not necessarily a metric) for comparing partition. Every partition $\mathrm{p} \in \mathbb{P}_S$ is assigned the vector $V_\mathrm{p} = (d(\mathrm{p},\mathrm{p}_1), d(\mathrm{p},\mathrm{p}_2), \ldots, d(\mathrm{p},\mathrm{p}_m))$ (feature map), and the kernel is accordingly given by $k_{(d,\mathcal{P})}(\mathrm{p},\mathrm{p}') = \langle V_\mathrm{p}, V_{\mathrm{p}'}\rangle$.

Justified in part by the intuition that the prototypes can be chosen by experts in the task at hand to stand for the different characteristic features of the data population, prototype-based measures have gained significant acknowledgment in the scientific community. Both empirical and theoretical evidence have been provided in support of their use in pattern recognition tasks, [**?**]. However, in the case of structured data, this approach falls short in capturing the essential facets of the intrinsic proximity notion.

To illustrate this fact, let us consider a simple example that can be easily generalized. The partitions of $S = \{x_1,x_2,x_3,x_4\}$ are described through the feature vectors listed below using the prototype set $\mathcal{P}$ consisting of the partitions $\mathrm{p}_1 = \{\{\{x_1,x_2,x_3,x_4\}, \mathrm{p}_2 = \{\{x_1,x_2,x_4\},\{x_3\}\}, \mathrm{p}_3 = \{\{x_1,x_2\},\{x_3,x_4\}\}, \mathrm{p}_4 = \{\{x_1,x_3\},\{x_2\},\{x_4\}\}$, and $\mathrm{p}_5 = \{\{x_1\},\{x_2,x_3,x_4\}\}\}$, and the β-entropy metric d_β ($\beta = 4$) on $\mathbb{P}_S$ defined by $d_\beta(\mathrm{p},\mathrm{p}') = \mathcal{H}_\beta(\mathrm{p}|\mathrm{p}') + \mathcal{H}_\beta(\mathrm{p}'|\mathrm{p})$ where the condition β-entropy is given by

$$\mathcal{H}_\beta(\mathrm{p}|\mathrm{p}') = \frac{1}{2^{1-\beta}-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{k'}\left(\frac{n_{ij}}{n}\right)^\beta - \sum_{j=1}^{k'}\left(\frac{n_j}{n}\right)^\beta\right).$$

$$\{\{x_1,x_2,x_3,x_4\}\} \overset{\varphi}{\mapsto} (0,0.8,1,1.1,0.8)$$
$$\{\{x_1,x_2,x_3\},\{x_4\}\} \overset{\varphi}{\mapsto} (0.8,0.6,0.3,0.3,0.6)$$
$$\{\{x_1,x_2,x_4\},\{x_3\}\} \overset{\varphi}{\mapsto} (0.8,0,0.3,0.4,0.8)$$
$$\{\{x_1,x_2\},\{x_3,x_4\}\} \overset{\varphi}{\mapsto} (1,0.3,0,0.2,0.3)$$
$$\{\{x_1,x_2\},\{x_3\},\{x_4\}\} \overset{\varphi}{\mapsto} (1.1,0.3,0.1,0.1,0.4)$$
$$\{\{x_1,x_3,x_4\},\{x_2\}\} \overset{\varphi}{\mapsto} (0.8,0.6,0.3,0.3,0.6)$$
$$\{\{x_1,x_3\},\{x_2,x_4\}\} \overset{\varphi}{\mapsto} (1,0.3,0.3,0.1,0.3)$$
$$\{\{x_1,x_3\},\{x_2\},\{x_4\}\} \overset{\varphi}{\mapsto} (1.1,0.4,0.2,0,0.4)$$
$$\{\{x_1,x_4\},\{x_2,x_3\}\} \overset{\varphi}{\mapsto} (1,0.3,0.3,0.2,0.3)$$
$$\{\{x_1\},\{x_2,x_3,x_4\}\} \overset{\varphi}{\mapsto} (0.8,0.6,0.3,0.4,0)$$
$$\{\{x_1\},\{x_2,x_3\},\{x_4\}\} \overset{\varphi}{\mapsto} (1.1,0.4,0.2,0.1,0.3)$$
$$\{\{x_1,x_4\},\{x_2\},\{x_3\}\} \overset{\varphi}{\mapsto} (1.1,0.3,0.2,0.1,0.4)$$
$$\{\{x_1\},\{x_2,x_4\},\{x_3\}\} \overset{\varphi}{\mapsto} (1.1,0.3,0.2,0.1,0.3)$$
$$\{\{x_1\},\{x_2\},\{x_3,x_4\}\} \overset{\varphi}{\mapsto} (0.1,0.4,0.1,0.1,0.3)$$
$$\{\{x_1\},\{x_2\},\{x_3\},\{x_4\}\} \overset{\varphi}{\mapsto} (1.1,0.3,0.1,0.1,0.3)$$

For $\mathrm{p} := \{\{x_1\},\{x_2\},\{x_3,x_4\}\}$, $\mathrm{p}' := \{\{x_1,x_2\},\{x_3,x_4\}\}$ and $\mathrm{p}'' := \{\{x_1,x_2,x_3,x_4\}\}$, we have that $\mathrm{p} \prec \mathrm{p}' \prec \mathrm{p}''$ and $k_{(d_\beta,\mathcal{P})}(\mathrm{p},\mathrm{p}') = 0.24$ while $k_{(d_\beta,\mathcal{P})}(\mathrm{p},\mathrm{p}'') = 0.68$, which contradicts Bottom-Up Collinearity. Moreover, if we consider $\mathrm{p} := \{\{x_1,x_2\},\{x_3\},\{x_4\}\}$, $\mathrm{p}' := \{\{x_1,x_2\},\{x_3,x_4\}\}$ and $\mathrm{p}'' := \{\{x_1,x_2,x_3,x_4\}\}$, then $k_{(d_\beta,\mathcal{P})}(\mathrm{p}',\mathrm{p}'') = 0.70$ and $k_{(d_\beta,\mathcal{P})}(\mathrm{p},\mathrm{p}'') = 0.78$, hence Top-Down Collinearity also fails.

As for Meet Predominance, take $\mathrm{p} := \{\{x_1\},\{x_2,x_3,x_4\}\}$ and $\mathrm{p}' := \{\{x_1,x_2\},\{x_3,x_4\}\}$. Then, $\mathrm{p} \wedge \mathrm{p}' = \{\{x_1\},\{x_2\},\{x_3,x_4\}\}$, and $k_{(d_\beta,\mathcal{P})}(\mathrm{p},\mathrm{p}') = 1.06$; however, $k_{(d_\beta,\mathcal{P})}(\mathrm{p}\wedge\mathrm{p}',\mathrm{p}) = 0.35$ and $k_{(d_\beta,\mathcal{P})}(\mathrm{p}\wedge\mathrm{p}',\mathrm{p}') = 0.39$.

5 Conclusions

In this paper, the almost unexplored class of positive kernels for comparing partitions was rigorously studied on the basis of the structural properties that govern the natural proximity between partitions. Several families of such kernels were introduced and their theoretical foundations have been explained. In particular, two generic procedures for designing hidden-feature-based kernels in compliance with the structural properties have been given, which can be straightforwardly adapted for the comparison of other structured data.

The main motivation of this paper lies in the new perspective that positive kernels provide in the scope of consensus methods. In this regard, an important future direction consists of the development of preimage methods that advance the performance of the heuristics used currently to approximate consensus solutions.

6 Appendix: Proofs of the results

PROOF (Proposition 1). *Suppose that the pairs in J have been listed. To each partition* p*, the feature map φ assigns a vector $\varphi(\mathrm{p})$ whose number of components is twice the number of pairs in J. Each pair $\{x, x'\} \in J$ is assigned two components of the vector $\varphi(\mathrm{p})$ denoted by $\varphi(\mathrm{p})(x, x', i)$, $i = 0, 1$, $\varphi(\mathrm{p})(x, x', i) = i \cdot \delta(x, x') \cdot \sqrt{\omega(\{x, x'\})} + (1 - i) \cdot (1 - \delta(x, x'))\sqrt{\nu(\{x, x'\})}$, where $\delta(x, x') = 1$ if $\{x, x'\} \in J(\mathrm{p})$, and $\delta(x, x') = 0$ otherwise. Only one of these components will be positive, the other is necessarily zero. Note also that, given any partitions* p *and* p'*, we have the following: (a) if a pair $\{x, x'\} \in N_{00}$, then $\varphi(\mathrm{p})(x, x', 0) = \varphi(\mathrm{p}')(x, x', 0) = 0$ and $\varphi(\mathrm{p})(x, x', 1) = \varphi(\mathrm{p}')(x, x', 1) = \sqrt{\omega(\{x, x'\})}$; (b) if a pair $\{x, x'\} \in N_{01}$, then $\varphi(\mathrm{p})(x, x', 0) = 0$, but $\varphi(\mathrm{p}')(x, x', 0) = \sqrt{\nu(\{x, x'\})}$, and $\varphi(\mathrm{p})(x, x', 1) = \sqrt{\omega(\{x, x'\})}$, but $\varphi(\mathrm{p}')(x, x', 1) = 0$; (c) if a pair $\{x, x'\} \in N_{10}$, then $\varphi(\mathrm{p})(x, x', 0) = \sqrt{\nu(\{x, x'\})}$, but $\varphi(\mathrm{p}')(x, x', 0) = 0$, and $\varphi(\mathrm{p})(x, x', 1) = 0$, but $\varphi(\mathrm{p}')(x, x', 1) = \sqrt{\omega(\{x, x'\})}$; and (d) if a pair $\{x, x'\} \in N_{11}$, then $\varphi(\mathrm{p})(x, x', 0) = \varphi(\mathrm{p}')(x, x', 0) = \sqrt{\nu(\{x, x'\})}$ and $\varphi(\mathrm{p})(x, x', 1) = \varphi(\mathrm{p}')(x, x', 1) = 0$. It is now straightforward to verify that $\langle \varphi(\mathrm{p}), \varphi(\mathrm{p}') \rangle$ equals to*

$$\sum_{\{x,x'\} \in J} \big[\varphi(\mathrm{p})(x, x', 0) \cdot \varphi(\mathrm{p}')(x, x', 0) + \varphi(\mathrm{p})(x, x', 1) \cdot \varphi(\mathrm{p}')(x, x', 1)\big]. \quad (8)$$

PROOF (Proposition 2) . *Let $\mathrm{p} \preceq \mathrm{p}' \preceq \mathrm{p}''$. Let N_{00} and N_{11} (resp. N'_{00} and N'_{11}) denote the sets of agreement between* p *and* p' *(resp.* p *and* p''*). Since* p *refines both* p' *and* p''*, $N_{00} = N'_{00}$. Since* p' *refines* p''*, $N'_{11} \subseteq N_{11}$. This proves that every term in the expansion of $k_{\omega\nu}(\mathrm{p}, \mathrm{p}'')$ is a term in the expansion of $k_{\omega\nu}(\mathrm{p}, \mathrm{p}')$. Therefore, $k_{\omega\nu}(\mathrm{p}, \mathrm{p}') \geq k_{\omega\nu}(\mathrm{p}, \mathrm{p}'')$.*

Under the same hypothesis, let N''_{00} and N''_{11} be the sets of agreement between p' *and* p''*. Then, $N'_{11} = N''_{11}$ and $N'_{00} \subseteq N''_{00}$. Thus, every term in the expansion of $k_{\omega\nu}(\mathrm{p}'', \mathrm{p})$ is a term in the expansion of $k_{\omega\nu}(\mathrm{p}'', \mathrm{p}')$, which forces $k_{\omega\nu}(\mathrm{p}'', \mathrm{p}') \geq k_{\omega\nu}(\mathrm{p}'', \mathrm{p})$.*

Since the clusters of $\mathrm{p} \wedge \mathrm{p}'$ are all the possible nonempty intersections between the clusters of p *and the clusters of* p'*, $N_{00} = N'''_{00}$ and $N_{11} \subseteq N'''_{11}$, where N'''_{00} and N'''_{11} denote the sets of agreements between* p *and $\mathrm{p} \wedge \mathrm{p}'$. Hence, $k_{\omega\nu}(\mathrm{p}, \mathrm{p} \wedge \mathrm{p}') \geq k_{\omega\nu}(\mathrm{p}, \mathrm{p}')$.*

PROOF (Theorem 1). *Let $\mathrm{p} \preceq \mathrm{p}' \preceq \mathrm{p}''$. Since $\mathrm{p} \preceq \mathrm{p}'$ and $\mathrm{p} \preceq \mathrm{p}''$, condition (ii) assures that if $f_i(\mathrm{p}) > 0$, then both $f_i(\mathrm{p}')$ and $f_i(\mathrm{p}'')$ are also positive. Thus, $f_i(\mathrm{p})f_i(\mathrm{p}')$ and $f_i(\mathrm{p})f_i(\mathrm{p}'')$ are positive if, and only if, $f_i(\mathrm{p})$ is positive. In particular, these products are both zero or both positive. Moreover, in the case that $f_i(\mathrm{p})f_i(\mathrm{p}')$ and $f_i(\mathrm{p})f_i(\mathrm{p}'')$ are positive, since $\mathrm{p}' \preceq \mathrm{p}''$ and $f_i(\mathrm{p}') > 0$, (ii) once again ensures that $f_i(\mathrm{p}') \geq f_i(\mathrm{p}'')$. Therefore, each positive term in the expansion of $k(\mathrm{p}, \mathrm{p}')$ is greater than the corresponding term in the expansion of $k(\mathrm{p}, \mathrm{p}'')$, which proves $k(\mathrm{p}, \mathrm{p}') \geq k(\mathrm{p}, \mathrm{p}'')$.*

On the other hand, as we have noted above, if $\mathrm{p} \preceq \mathrm{p}''$, then $k(\mathrm{p}, \mathrm{p}'') = \sum_i f_i(\mathrm{p}'')f_i(\mathrm{p}) = \sum_{i, f_i(\mathrm{p})>0} f_i(\mathrm{p}'')f_i(\mathrm{p})$. By virtue of condition (iii), $k(\mathrm{p}'', \mathrm{p}) = \sum_{i, f_i(\mathrm{p})>0} f_i(\mathrm{p}'')f_i(\mathrm{p}) \leq \sum_{i, f_i(\mathrm{p})>0} f_i(\mathrm{p}'')f_i(\mathrm{p}') + \sum_{i, f_i(\mathrm{p})=0} f_i(\mathrm{p}'')f_i(\mathrm{p}') = k(\mathrm{p}'', \mathrm{p}')$.

Finally, by virtue of condition (iv) we can deduce that if $f_i(\mathrm{p})f_i(\mathrm{p}')$ is positive term in the expansion of $k(\mathrm{p}, \mathrm{p}')$, then the corresponding term $f_i(\mathrm{p})f_i(\mathrm{p} \wedge \mathrm{p}')$ in the expansion of $k(\mathrm{p}, \mathrm{p} \wedge \mathrm{p}')$ is also positive. Besides, since $\mathrm{p} \wedge \mathrm{p}' \preceq \mathrm{p}$, (ii) guarantees $f_i(\mathrm{p} \wedge \mathrm{p}') > f_i(\mathrm{p}')$, and consequently, $k(\mathrm{p}, \mathrm{p} \wedge \mathrm{p}') \geq k(\mathrm{p}, \mathrm{p}')$.

PROOF (Proposition 3) . *Let us consider any two partitions $\mathrm{p} = \{\mathrm{c}_1, \mathrm{c}_2, \ldots, \mathrm{c}_s\}$ and $\mathrm{p}' = \{\mathrm{c}'_1, \mathrm{c}'_2, \ldots, \mathrm{c}'_r\}$ of S, with n_i, n'_j, and n_{ij} denoting the sizes of ith cluster c_i of* p*, the jth cluster c'_j of* p'*, and the intersection $\mathrm{c}_i \cap \mathrm{c}'_j$.*

$$\begin{aligned}
k_\mu(\mathrm{p}, \mathrm{p}') &= \sum_i f_i(\mathrm{p})f_i(\mathrm{p}') \\
&= \sum_{i=1}^{s}\sum_{j=1}^{r}\left(\sum_{\mathfrak{s} \subseteq \mathrm{c}_i \cap \mathrm{c}'_j} \frac{|\mathfrak{s}|}{|\mathrm{c}_i|} \cdot \frac{|\mathfrak{s}|}{|\mathrm{c}'_j|}\right) \\
&= \sum_{i=1}^{s}\sum_{j=1}^{r}\left(\sum_{u=1}^{n_{ij}} \binom{n_{ij}}{u} \cdot \frac{u}{n_i} \cdot \frac{u}{n'_j}\right) \\
&= \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{1}{n_i n'_j} \cdot \left(\sum_{u=1}^{n_{ij}} \binom{n_{ij}}{u} \cdot u^2\right). \quad (9)
\end{aligned}$$

On the other hand, note that

$$\begin{aligned}
\sum_{u=1}^{n_{ij}} \binom{n_{ij}}{u} \cdot u^2 &= \sum_{u=1}^{n_{ij}} \frac{n_{ij}!}{u!(n_{ij} - u)!} \cdot u^2 \\
&= \sum_{u=1}^{n_{ij}} \frac{(n_{ij} - 1)!}{(u - 1)!(n_{ij} - u)!} \cdot n_{ij} \cdot u \\
&= n_{ij} \sum_{u=1}^{n_{ij}} \binom{n_{ij} - 1}{u - 1} \cdot u. \quad (10)
\end{aligned}$$

In addition, since

$$\binom{n_{ij}-1}{u-1}\cdot u = \binom{n_{ij}-1}{u-1}\cdot(u-1)+\binom{n_{ij}-1}{u-1} = (n_{ij}-1)\binom{n_{ij}-2}{u-2}+\binom{n_{ij}-1}{u-1}$$

Thus,

$$\sum_{u=1}^{n_{ij}}\binom{n_{ij}-1}{u-1}\cdot u = (n_{ij}-1)\sum_{u=1}^{n_{ij}}\binom{n_{ij}-2}{u-2}+ \sum_{u=1}^{n_{ij}}\binom{n_{ij}-1}{u-1},$$

which gives us

$$\sum_{u=1}^{n_{ij}}\binom{n_{ij}-1}{u-1}\cdot u = (n_{ij}-1)\cdot 2^{n_{ij}-2}+2^{n_{ij}-1} = 2^{n_{ij}-2}\cdot(n_{ij}+1). \tag{11}$$

Substituting (11) in (10), we get

$$\sum_{u=1}^{n_{ij}}\binom{n_{ij}}{u}\cdot u^2 = 2^{n_{ij}-2}\cdot(n_{ij}^2+n_{ij}),$$

and then replacing (10) in (9), we conclude

$$k_\mu(\mathrm{p},\mathrm{p}') = \frac{1}{4}\sum_{i=1}^{s}\sum_{j=1}^{r}2^{n_{ij}}\cdot\left(\frac{n_{ij}^2+n_{ij}}{n_i n_j'}\right).$$

To prove that k_μ fulfills conditions $(i)-(iv)$ in Theorem 1, note that $\mu(\mathfrak{s},\mathrm{p})$ is either zero or the ratio between the respective sizes of two sets and therefore nonnegative. Furthermore, if $\mathrm{p}\preceq\mathrm{p}'$, then every cluster of p is contained in some cluster of p'. If a subset $\mathfrak{s}$ of S is contained into a cluster c of p, it will be also contained into the cluster c' of p' that contains c. This is, $\mathfrak{s}\subseteq\mathrm{c}\subseteq\mathrm{c}'$. Hence, the ratio between the size of $\mathfrak{s}$ and the size of c is greater than the ratio between the size of $\mathfrak{s}$ and the size of c'. Thus, if $\mu(\mathfrak{s},\mathrm{p})>0$, then $\mu(\mathfrak{s},\mathrm{p}')>0$ and $\mu(\mathfrak{s},\mathrm{p})\geq\mu(\mathfrak{s},\mathrm{p}')$. This proves (ii). In order to verify (iii) (still assuming $\mathrm{p}\preceq\mathrm{p}'$), note that, if $\mu(\mathfrak{s},\mathrm{p})>\mu(\mathfrak{s},\mathrm{p}')$, then $\mathfrak{s}\subseteq\mathrm{c}\subset\mathrm{c}'$. If $\mathfrak{s}$ does not cover the entire c, consider the subset $\mathfrak{s}'=\mathfrak{s}\cup(\mathrm{c}'-\mathrm{c})$. Obviously, $\mathfrak{s}'\subset\mathrm{c}'$; however, there is no cluster of p containing $\mathfrak{s}'$. Thus, $\mu(\mathfrak{s}',\mathrm{p}')>0$, but $\mu(\mathfrak{s}',\mathrm{p})=0$. Moreover, given that $\frac{|\mathfrak{s}|}{|\mathrm{c}|}<1$

$$\begin{aligned}\mu(\mathfrak{s},\mathrm{p})-\mu(\mathfrak{s},\mathrm{p}') &= \frac{|\mathfrak{s}|}{|\mathrm{c}|}-\frac{|\mathfrak{s}|}{|\mathrm{c}'|}\\ &= \frac{|\mathfrak{s}|}{|\mathrm{c}|}\cdot\left(\frac{|\mathrm{c}'|-|\mathrm{c}|}{|\mathrm{c}'|}\right)\\ &< \frac{|\mathrm{c}'|-|\mathrm{c}|}{|\mathrm{c}'|}\\ &< \frac{|\mathrm{c}'|-|\mathrm{c}|}{|\mathrm{c}'|}+\frac{|\mathfrak{s}|}{|\mathrm{c}'|} \qquad (12)\\ &= \mu(\mathfrak{s}',\mathrm{p}')\end{aligned}$$

If $\mathfrak{s}=\mathrm{c}$, then $\mu(\mathfrak{s},\mathrm{c})=1$. In this case, if $\mathrm{c}'-\mathrm{c}$ is another (single) cluster of p, say c_0, then

$$\mu(\mathrm{c},\mathrm{p})-\mu(\mathrm{c},\mathrm{p}')+\mu(\mathrm{c}_0,\mathrm{p})-\mu(\mathrm{c}_0,\mathrm{p}')=1=\mu(\mathrm{c}',\mathrm{p}');$$

otherwise, for $\mathfrak{s}'=\mathrm{c}'-\mathrm{c}$, $\mu(\mathfrak{s}',\mathrm{p})=0$ because it is the union of at least two clusters of p, and

$$\mu(\mathrm{c},\mathrm{p})-\mu(\mathrm{c},\mathrm{p}')=1-\frac{|\mathrm{c}|}{|\mathrm{c}'|}=\frac{|\mathrm{c}'|-|\mathrm{c}|}{|\mathrm{c}'|}=\mu(\mathfrak{s}',\mathrm{p}').$$

Lastly, observe that for any partitions p and p', if the subset $\mathfrak{s}$ is such that both $\mu(\mathfrak{s},\mathrm{p})$ and $\mu(\mathfrak{s},\mathrm{p}')$ are positive, then there are clusters c in p and c' in p' with $\mathfrak{s}\subseteq\mathrm{c}$ and $\mathfrak{s}\subseteq\mathrm{c}'$. This implies $\mathfrak{s}\subseteq(\mathrm{c}\cap\mathrm{c}')$. But $\mathrm{c}\cap\mathrm{c}'$ is a cluster of $\mathrm{p}\wedge\mathrm{p}'$, and hence $\mu(\mathfrak{s},\mathrm{p}\wedge\mathrm{p}')$ is positive.

PROOF (Proposition 4). *By definition, the similarity measure σ has been considered to be non-negative. Therefore, all the components of the vector $\varphi_\sigma(\mathrm{p})$ are non-negative. This gives (i).*

Let us suppose now that $\mathrm{p}\prec\mathrm{p}'$. In this case, every cluster of p is contained in some cluster of p', and vice versa. Thus, if $\mathfrak{s}_i$ is a subset of S containing some cluster c' of p', then $\mathfrak{s}_i$ contains at least one cluster c of p. Moreover, every singleton cluster of p' is a singleton cluster of p. Accordingly, $c_i(\mathrm{p})\geq c_i(\mathrm{p}')$ and $\nu_i(\mathrm{p}')=0$ forces $\nu_i(\mathrm{p})=0$. Hence, $f_i(\mathrm{p}')=1+\frac{\nu_i(\mathrm{p}')\cdot diam(\mathfrak{s}_i)}{c_i(\mathrm{p}')}\geq 1+\frac{\nu_i(\mathrm{p})\cdot diam(\mathfrak{s}_i)}{c_i(\mathrm{p})}=f_i(\mathrm{p})$. This gives (v).

Lastly, under the same assumption of $\mathrm{p}\prec\mathrm{p}'$, we shall verify (vi). In this regard, note that

$$\sum_{i,c_i(\mathrm{p}')>0}f_i(\mathrm{p}')-\sum_{i,c_i(\mathrm{p}')>0}f_i(\mathrm{p}) = \sum_{\substack{i,c_i(\mathrm{p}')>0\\ \nu_i(\mathrm{p}')=1}} diam(\mathfrak{s}_i)\cdot\left(\frac{\nu_i(\mathrm{p}')}{c_i(\mathrm{p}')}-\frac{\nu_i(\mathrm{p})}{c_i(\mathrm{p})}\right).$$

If $\nu_i(\mathrm{p}')=1$ and $c_i(\mathrm{p})>c_i(\mathrm{p}')>0$, then we shall consider the subset $\mathfrak{s}_{j_i}$ obtained from $\mathfrak{s}_i$ by applying the following procedure:

Erase one element from each cluster c' of p' *contained in* $\mathfrak{s}_i$. The removal should be done in such a way that one cluster c of p that is not a cluster of p' remains invariant.

Thus, we get that $c_{j_i}(\mathrm{p}') = 0$ *and* $c_{j_i}(\mathrm{p}) > 0$. *In addition, since* $diam(\mathfrak{s}_i) \leq 1$, $diam(\mathfrak{s}_i)\left(\frac{\nu_i(\mathrm{p}')}{c_i(\mathrm{p}')} - \frac{\nu_i(\mathrm{p})}{c_i(\mathrm{p})}\right) < 1 \leq \left(1 + \frac{\nu_{i_j}(\mathrm{p})\cdot diam(\mathfrak{s}_{j_i})}{c_{j_i}(\mathrm{p})}\right)$. *This proves*

$$\sum_{i, c_i(\mathrm{p}')>0} f_i(\mathrm{p}') - \sum_{i, c_i(\mathrm{p}')>0} f_i(\mathrm{p}) \leq \sum_{i, c_i(\mathrm{p}')=0} f_i(\mathrm{p}).$$

and consequently (vi).

References:

[1] Kakeru Narita, Teruhisa Hochin, Yoshihiro Hayashi, Hiroki Nomiya: Incremental Hierarchical Clustering for Data Insertion and Its Evaluation *International Journal of Software Innovation (IJSI)*, 8(2), 2022, 1-22.

[2] Pasi Fränti, Sami Sieranoja, How much can k-means be improved by using better initialization and repeats?, Pattern Recognition, Vol. 93, 2019, 95-112.

[3] Seyed Saeed Hamidi, Ebrahim Akbari, Homayun Motameni, Consensus clustering algorithm based on the automatic partitioning similarity graph, Data & Knowledge Engineering, Vol. 124, 2019, 101754.

[4] Jon M. Kleinberg, An Impossibility Theorem for Clustering, in *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, MIT Press, (2002) 446–453.

[5] Margareta Ackerman, Shai Ben-David, Simina Brânzei, David Loker: Weighted clustering: Towards solving the user's dilemma. *Pattern Recognit.* 120, 2021, 108152.

[6] Carlsson, Gunnar and Mamoli, Facundo, Classifying Clustering Schemes, *Foundations of Computational Mathematics*, 13 (2), (2013) 221–252.

[7] Jyrko Correa-Morris, An indication of unification for different clustering approaches, *Pattern Recognit.*, 46 (9), (2013) 2548–2561.

[8] Jean-Pierre Barthélemy and Bruno Leclerc, The Median Procedure for Partitions, *Partitioning Data Sets, American Mathematics Society, Series in Discrete Math*, 1995, 3-34.

[9] Marina Meilă, Local equivalences of distances between clusterings - a geometric perspective. Mach. Learn. 86(3), 2012, 369-389.

[10] Jyrko Correa-Morris, Abel Urra-Yglesias, Estefano Reyes, Juan Martínez, Belarmino Gonzalez: Comparing Partitions: Metric Characterizations, Mean Partition, and Distortion, SAI (1), 2021, 857-875.

[11] Jyrko Correa-Morris, Comparing Partitions: Shortest Path Length Metrics and Submodularity, *Intern. J. of Math. Models and Meth. in Appl. Sci.*, 13, 2019, 45-51.

[12] Serratosa, F. Redefining the Graph Edit Distance. SN COMPUT. SCI. 2, 2021, 438.

[13] Shawe-Taylor, J., and Cristianini, N., Kernel Methods for Pattern Analysis, Cambridge: Cambridge University Press, 2004.

[14] Marco Cuturi, Positive definite kernels in machine learning, arXiv preprint arXiv:0911.5367, 2009 - arxiv.org.

[15] Bo Ma, Hui-yang Qu, Hau-san Wong, Kernel clustering-based discriminant analysis, *Pattern Recognit.*, 40(1), 2007, 324-327.

[16] Inderjit S. Dhillon and Yugiang Guan and Brian Kulis, Kernel k-means: spectral clustering and normalized cuts, *In KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, 551-556.

[17] Zhu Y and Ting K., Kernel-based clustering via Isolation Distributional Kernel. *Information Systems*, 117, 2023, 102212.

[18] Yang X, Lin G, Liu Y, Nie F and Lin L., Fast Spectral Embedded Clustering Based on Structured Graph Learning for Large-Scale Hyperspectral Image, *IEEE Geoscience and Remote Sensing Letters*, 19, 2020, 1-5.

[19] Wang R, Lu J, Lu Y, Nie F and Li X, Discrete and Parameter-Free Multiple Kernel k - Means, *IEEE Transactions on Image Processing*, 31, 2022, 2796-2808.

[20] Liu H, Chen J, Dy J and Fu Y., Transforming Complex Problems into K-means Solutions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7), 2023, 1-20.

[21] Brijnesh J. Jain, The Mean Partition Theorem in consensus clustering, *Pattern Recognition*, 79, 2018, 427-439.

[22] Sandro Vega-Pons, Jyrko Correa-Morris, José Ruiz-Shulcloper: Weighted partition consensus via kernels. Pattern Recognit. 43(8), 2010, 2712-2724.

[23] William M. Rand, Objective Criteria for the Evaluation of Clustering Methods, emphJournal of the American Statistical Association, 66(336), 1971, 668-675.

[24] Sandro Vega-Pons, Xiaoyi Jiang, José Ruiz-Shulcloper: Segmentation Ensemble via Kernels, ACPR 2011, 2011, 686-690.

[25] Lucas Franek, Xiaoyi Jiang: Ensemble clustering by means of clustering embedding in vector spaces. Pattern Recognit. 47(2), 2014, 833-842.

[26] Jyrko Correa-Morris: The median partition and submodularity, *Appl. Math. Comput.* 410, 2021, 126450.

Contribution of Individual Authors to the Creation of a Scientific Article

Jyrko Correa-Morris was responsible for the sections of the paper.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

Partial support for this research has been provided by the U.S. Department of Education under grant P120A200005.

Conflicts of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.